Marco Frasca
Curriculum vitae

Contents

Personal data

SURNAME	Frasca
NAME	Marco
E-MAIL	marco.frasca at unimi dot it
WEB-SITE	http://frasca.di.unimi.it

1 Short Biography

1.1 Current Job

Associate Professor
Department of Computer Science, University of Milan, Italy

1.2 Work Experiences

- January 2023 - Today. Full-time Associate Professor at Department of Computer Science, University of Milan.
- May 2017 - Decembre 2022. 3 years position as Assistant Professor at Department of Computer Science, University of Milan.
- April 2013 - March 2017. 4 years post-doc research grant at Department of Computer Science, University of Milan, project title "Graph based neural network algorithms for the analysis of biological networks"
- June 2012 - March 2013. 1 year post-doc research grant at Department of Life Sciences of University of Milan, project title "Development of algorithms and software tools for the analysis of chromatin role in eukaryotic genome"
- January 2009 - December 2011. 3 years Ph.D. in Computer Science with grant at Department of Computer Science, University of Milan
- July 2008 - October 2008. Internship at Metoda S.P.A., Salerno, Italy, for the study and development of security protocols, certification and mutual authentication within SAFE project: "Sistema di Anamnesi in Fase di Emergenza".
- January 2007- December 2007. 1 year grant for promising teachers for high school at University of Salerno, Italy.
- January 2006 - September 2006. Employer of Texa S.P.A. company, Parma, Italy, with endless contract position as software programmer.

1.3 Education

- January 2009 - December 2011. Ph.D. student in the Ph.D. course in Computer Science, with 3 years grant, at Department of Computer Science, University of Milan. Title of Ph.D. thesis: "Graph-based approaches for imbalanced data in functional genomics".
  Supervisor: Prof. Alberto Bertoni
  Co-supervisor: Prof. Giorgio Valentini

- October 2006 - May 2008. Qualification course, with 1 year grant, for teaching mathematics at high school, gained at University of Salerno, Italy. Final evaluation: 80/80.
- May 2005. Master degree (five years course) in Computer Science at University of Salerno, Italy. Thesis title: "L'operazione di composizione nella famiglia dei codici massimali prefissi". Supervisor: Prof.ssa Clelia De Felice.
  Final evaluation: 110/110 cum laude.

1.3.1 Research internships abroad and participation in international summer schools

Research internships abroad

- September 2017. Research visit at the Department of Medicine, Division of Neurology, Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto, to work on the disease-gene prioritization for the Alzheimer's disease (AD) with Prof. Ming Zhang, exploiting the novel approach to embed in the computational core information coming from diseases sharing genetic characteristics with AD.
- September – October 2016. Research internships at the Institute of Molecular Biology (IMB) of Johannes Gutenberg University of Mainz, for a collaboration with the Computational Biology and Data Mining Group, directed by Prof. Miguel Andrade Navarro, to work on the design and development of computational methodologies to find the most informative data sources for specific human genetic disorders, with the end of determining their causal genes.
- September – October 2014. Visiting Researcher at the Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada, to collaborate with Prof. Quaid Morris, head of laboratory, on the analysis of computational methodologies to eliminate hierarchical inconsistencies of label biases in label propagation algorithms when classes are structured as directed acyclic graphs, in the context of Life Sciences.

Attendance at international summer schools

- 14-19 August 2011. "From Data to Models in Biological Systems", Kandersteg (Switzerland). Organized by the Swiss Institute of Bioinformatics (SIB).
- 12-19 June 2010. "Statistical and Machine Learning Methods in Computational Biology", Lipari (Italy). Organized by the Jacob T. Schwartz International School for Scientific Research.

1.3.2 Collaborations with national and international research groups

The candidate has collaborated and collaborates with several national and international research groups and institutions, including the Royal Holloway University of London, the University Charité of Berlin, the Johannes Gutenberg University of Mainz (Germany), the Fondazione Centro San Raffaele in Milan (Italy), and the National Institute for Cancers of Milan. In the following a detailed list.

Collaborations with international research groups:

- Quantum Blockchain Technologies, as Machine Learning expert to develop ML intelligence for cryptocurrency context.
- Past collaborations with the Royal Holloway University of London, with the group of Prof. Alberto Paccanaro, involved the development of a web tool for the implementation of semantic similarity measures between terms of hierarchies structured as directed acyclic graphs [27,

47]. Ongoing collaborations regard the study of patient networks for modelling the phenotype predictions from biomolecular profiles [12], and the problem of associating human complex diseases with their causative genes.

- Participation as member of the AnacletoLab laboraroty of University of Milan at the international challenge CAFA3 (Critical Assessment of Functional Annotation) within the Special Interest Group "Protein Function Prediction" of ISCB (International Society of Computational Biology) [13]. The Special Interest Group gathers the main international research groups for the automated protein function prediction[1].
- Collaboration with the Computational Biology and Data Mining Group of Johannes Gutenberg University of Mainz (Germany), directed by Prof. Miguel Andrade Navarro, to study and develop computational methodologies to discover novel causal genes for human genetic disorders [38].
- ECLT (European Center for Living Technologies) of Venice (Italy) for the study of methods based on the game theory for the inference in biomolecular networks, with applications in the fields of Network Medicine and Molecular Biology [18].
- Group of Prof. Peter N. Robinson, head of the Computational Biology Group of the Max Planck Institute for Molecular Genetics, University Charité di Berlin, for studying algorithms for the hierarchical correction of predictions in biological taxonomies [45].

Collaborations with national research groups and research institutes:

- PRIN Project 2017, titled:"Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond", collaboration with Università di Pisa, Palermo and Piemonte Orientale to develop a new family of multi-criteria data structures, based on "learned" data stuctures. Actually, The following results are jointly achieved [31].
- Active. Collaboration with the Fondazione Centro San Raffaele[2] of Milan, with the group of doctor Federica Esposito, to learn predictors about the response of patients afflicted by multiple sclerosis to the treatment with FINGOLIMOD drug [2].
- National Council of Research (CNR), analysis of genetic diseases and the detection of new biomarkers associated with them [4, 5, 6].

1.3.3 Seminars held about his research activity

- "Neural algorithms based on graphs for the analysis of biological networks", Department of Computer Science, University of Milan, 09 April 2015.
- "Machine Learning for Bioinformatics and Personalized Medicine: a survey of my research activity at the Computer Science Dept UNIMI", Department of Computer Science, University of Milan, 20 April 2017.

1.3.4 Awards and Acknowledgments

- From April 2013. 4 years research grant as post-doc research fellow at the Department of Computer Science, University of Milan, Italy.
- June 2012 - March 2013. 1 year research grant as post-doc research fellow at the Department of Life Sciences of University of Milan.

[1] `http://biofunctionprediction.org/`
[2] `http://www.fondazionesanraffaele.it/`

- January 2009 - December 2011. 3 years research grant from the Italian national funding for Ph.D. courses at the Department of Computer Science, University of Milan.
- January 2007- December 2007. 1 year grant for promising teachers of mathematics at high school for the attendance of the two years course for teaching qualification at University of Salerno, Italy.

1.3.5 Participation as speaker in national and international conferences

- 2022 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM), online, 3-5 febbraio 2022. Presentazione del lavoro [31].
- 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Wurzburg, Germania, 16-20 settembre 2019. Presentazione del lavoro [34].
- 2019 9th 9th International Conference on Biomedical Engineering and Technology (ICBET 2019) March 28-30, 2019, Tokyo, Japan. Presentazione del lavoro [37].
- 2016 15th International Workshop on Data Mining in Bioinformatics (BIOKDD '16), within the ACM SIGKDD 2016 Conference on Knowledge Discovery and Data Mining,13-17 August, San Francisco, USA, to present the work [42].
- 2016 International Work-Conference on Bionformatics and Biomedical Engineering (IWBBIO 2016), Granada, Spain, 20-22 April 2016, to present the work [43].
- 2013 International Joint Conference on Neural Networks (IJCNN 2013), Dallas, Texas, 4-9 August 2013, to present the work [48].
- 2012 22th Italian Workshop on Neural Networks, Vietri sul Mare, Salerno, Italy, 17-19 May 2012, to present the work [49].
- 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Atens, Greece, 5-9 September 2011, to present the work [50].
- 2011 Eighth Annual Meeting of the Bioinformatics Italian Society, Pisa, ItalY, 20-22 June 2011, to present the work [51].
- 2010 20th Italian Workshop on Neural Networks, Vietri sul Mare, Salerno, Italy, 27-29 MAy 2010,to present the work [52].

2 Reviewer and Conference Organization Activities

The candidate is reviewer for several peer-reviewed national and international journal and conferences. The detailed list as follows:

International journals:

- Expert Systems with Application
- Neural Networks
- Neurocomputing
- IEEE Transactions on Neural Networks and Learning Systems

- BMC Bioinformatics
- Information Sciences

International conferences

- World Congress on Intelligent Control and Automation (WCICA 2016).
- International Joint Conference on Artificial Intelligence (IJCAI) 2015
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2014.
- International Conference on Artificial Neural Networks (ICANN), 2014.
- International Conference on Artificial Neural Networks (ICANN), 2013.

National conferences

- Italian Workshop on Neural Networks, WIRN 2015.
- Italian Workshop on Neural Networks, WIRN 2014.
- Italian Workshop on Neural Networks, WIRN 2012.

Moreover, the candidate is project reviewer for:

- the *Natural Sciences and Engineering Research Council of Canada* (NSERC), in the area *Discovery Grant proposal*[3].
- the Innovational Research Incentives Scheme (VIDI programme), for projects presented by young researchers (3–8 years from Ph.D.) of the Netherlands Organisation for Scientific Research (NWO, the Dutch Research Council)[4].

Member of the Program Committee of the following international workshops/conferences:

- International Conference on Artificial Neural Networks (ICANN 2020), Bratislava, Slovakia, settembre 15–18, 2020
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2019), October 13-16, 2019 Pittsburgh, PA, USA
- International Conference on Artificial Neural Networks (ICANN 2019) Munich, Germany,17–19 September 2019
- International Conference on Artificial Neural Networks (ICANN 2019) Munich, Germany,17–19 Semptember 2019`https://e-nns.org/icann2019/`
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018) Calgary, Alberta, Canada,15–20 April 2018[5]

[3] `http://www.nserc-crsng.gc.ca/index_eng.asp`
[4] `http://www.nwo.nl/en/funding/our-funding-instruments/nwo/innovational-research-incentives-scheme/vidi/index.html`
[5] `https://2018.ieeeicassp.org/Committee.asp`

- First International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2019), Barcelona, Spain, 20-22 March 2019[6]
- Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2018), Caparica, Portugal, September 6-8, 2018[7]
- International Conference on Signal Processing and Machine Learning (SPML 2018), Shanghai, China, November 28-30, 2018[8]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2018), Aalborg, Denmark, September 17-20, 2018[9]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2017), Tokyo, Japan, September 25-28, 2017[10]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2016), Vietri sul Mare, 13–16 settembre 2016[11].

He has been Guest Editor of the journal track of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2019-2022.

He has been Associate Editor of the International Journal of Neural Networks (ISSN:2249-2763 print version,E-ISSN: 2249-2771 electronic version, DOI:10.9735/2249-2763) from 2013 to 2017.

He is member of the Bioinformatics Italian Society (BITS) since 2017.

He has been member of the International Neural Network Society (INNS) form May 2013 to December 2014.

Finally, he has been session chair of the International Joint Conference of Neural Network (IJCNN) in 2013[12], in the session dedicated to Bioinformatics.

3 Teaching Activity and Student Tutoring

3.1 Teaching activity

- A.A 2019-today. Course of Laboratory of Algorithms and Data Structures (48 hours). Department of Computer Science, University of Milan.
- A.A. 2017-2019. Course of Laboratory of Databases (48 hours). Department of Computer Science, University of Milan, Italy.
- Year 2018. Laboratory of Data Mining and Machine Learning (25 ore). Second level Master in Data science for economics, business and finance, University of Milan, Italy.
- A.A. 2015-2016. Course of Laboratory of Algorithms (44 hours). Department of Mathematics, University of Milan, Italy.

[6] https://drive.google.com/file/d/1hLl_rXea028z6q9GJIxK5UK0LGrmJmR7/view
[7] https://eventos.fct.unl.pt/cibb2018/pages/organization
[8] http://www.spml.net
[9] http://mlsp2018.conwiz.dk/home.htm
[10] https://signalprocessingsociety.org/blog/mlsp-2017-2017-ieee-international-workshop-machine-learning-signal-processing
[11] http://www.conwiz.dk/cgi-all/mlsp2016/list_pc.pl
[12] http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=31079

- A.A. 2015-2016. Laboratory exercises for the Statistic and Data Analysis course, (4 hours). Department of Computer Science, University of Milan.
- A.A. 2015-2016. Third part of the Laboratory of Algorithms and Data Structures course (12 hours). Department of Computer Science, University of Milan.
- A.A. 2011-2015. Course of Laboratory of Algorithms (44 hours). Department of Mathematics, University of Milan.
- A.A. 2013-2014. Course of Informatics for biotechnology (50 hours). Department of Life Sciences, University of Milan.
- A.A 2010-2011. Course of Laboratory of Algorithms and Data Structures, evening session (48 hours). Department of Computer Science, University of Milan.

3.2 Co-supervision of bachelor and master thesis

At the Department of Computer Science, University of Milan, the candidate has been supervisor and co-supervisor of around 3 bachelor and master thesis per year.

4 Participation in Research Projects

He is/has been responsible of the following projects:

- Pfizer Global NASH ASPIRE Competitive Grant Program.

  `Title`. Optimizing Management of Comorbid NASH through a Multidisciplinary lntegration and Artificial lntelligence Alliance (OPTIMA-NASH).

  `Funder`. Pfizer.

  `Role`. Co-leader of Computer Science unit.

  `Coordinator University`. University of Milan (UNIMI).

  `Duration`: 2 anni, since 01/10/2022.

- PRIN PROJECT 2017.

  `Title: Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond.`

  `Funding institution: Italian Ministry of University (MIUR). Duration: 3 years, since 01/07/2019.`

  `Role: responsible for the UNIMI branch (Dipartimento di Informatica 'Giovanni degli Antoni').`

  `National Coordinator: Università di Pisa.`

- `Title: Hierarchical classification algorithms in biomedical taxonomies,`

  `Piano di Sostegno alla Ricerca 2015-2017`

  `Starting date: 01.06.19`

  `Duration: 1 year`

  `Personnel involved: 2`

- Title: Machine learning algorithms to handle label imbalance in biomedical taxonomies,
  Piano di Sostegno alla Ricerca 2015-2017
  Starting date: 01.01.18
  Duration: 1 year

The candidate participated in the following research projects:

- Title: "A predictive model or response to FINGOLIMOD: integration of clinics, neuroradiology and genetics". Financing body: Fondazione Centro San Raffaele, Duration: May-September 2016
- TItle: "Discovering Patterns in Multi-Dimensional Data". Year 2016.
  Coordinator: Prof. Goffredo Haus. Financing body: University of Milan
- Title: "Graph-based methodologies for the automated inference in bio-medical ontologies". Year 2016. Coordinator: Prof. Simone Bassis. Financing body: University of Milan
- Flagship project EPIGEN, May 2012, March 2013. Research project title: "Sviluppo di algoritmi e software per l'analisi dello stato della cromatina in genomi di eucarioti". Financing body: Consiglio Nazionale delle Ricerche (CNR)

5 Publications and Bibliometric Indicators

In the following the main bibliometric indicators:

- GOOGLE SCHOLAR (01.23)
  (https://scholar.google.it/citations?user=eduk3B8AAAAJ&hl=en)
  Citations: 803
  H-index: 13
- ELSEVIER SCOPUS (www.scopus.com, 01.23).
  Author ID: 36835331200
  Citations: 453.
  H-index: 11

6 Research Activity

The candidate's research activity mainly focused on the design and analysis of new machine learning methods, with applications in bioinformatics and computational biology and medicine. His Computer Science background, and the strong interest for Life Sciences branches fostered searching novel research lines as bridge between these two main areas. He contributed to consolidate the application of Hopfield networks to classification and ranking problems with the development of new models of parametric Hopfield networks, whose implementations are available at public software repositories.

In particular, the candidate developed methods suitable for classification problems characterized by unbalanced data, that is where one class is the large minority, like it happens in many real world problems in computational biology and medicine, like the gene function prediction, drug repositioning, gene-disease prioritization, human abnormal phenotypes prediction problems.

Furthermore, he devoted his research also to the study of methodologies suitable for classification problems where classes are structured as hierarchies with parent-child constraints. Namely, he studied multi-task classifiers, that is classifiers which predict at the same time multiple classes

for the same problem, and learn faster and more accurately one class by exploiting the learning of the other ones. Recent works indeed focused on the development on multi-task algorithms based on label propagation on labeled graphs.

Within the candidate's research activity it is possible to distinguish two main areas, machine learning and bioinformatics, each further divided into subareas.

6.1 Machine learning

In machine learning area we distinguish the following sub-areas:

6.1.1 Development of binary classifiers for highly unbalanced data through parametric Hopfield networks

This area regards classification problems modeled as semi-supervised label prediction in partially labeled graphs, where nodes represent the entities to be classified and edges the similarities among nodes, whereas a labeling is known just for a subset of nodes. The goal is to extend the labeling to the whole graph. In particular the focus is on problems where node labels are unbalanced, that is one class (usually the positive one) is highly under-represented that the other one, since several real-world problems posses this characteristic. The labeling imbalance must be handle to avoid a noticeable decay in the classification performance.

To this end, he developed a new model of parametric Hopfield network, *COSNet* (**CO**st-**S**ensitive Neural **Net**work), whose parameters are the neuron activation values and thresholds. The model conceptually separates the neuron labels and the neurons activation values, allowing the learning of activation values that counterbalance the predominance of negative neurons to handle the label imbalance. Positive neurons can thereby propagate their labels during the network dynamics avoiding trivial equilibrium points. *COSNet* has been compared with the state-of-the-art algorithms in complex real-world problems, such as the gene function prediction problem [28, 50]. A regularized version of the model has been proposed to manage some limit cases of the learning procedure of *COSNet* [29]. Moreover, the model is able also in inferring node ranking with regard to the class to be predicted, so as to rank higher nodes more likely to belong the class predicted [26, 48].

In order to exploit the potential of the novel graphical devices, such a model has been then re-designed to run in parallel on GPU architectures. The key strategy to speed up the computations is to partition nodes into independent sets so as to process each set in parallel by exploiting the power of GPU accelerators, while keeping asynchronous the overall dynamics, condition necessary to ensure the dynamics convergence. Furthermore, by adopting a sparse representation of both network connections and node labeling, leveraging the sparsity of input graphs and the scarcity of positive proteins, the method, named ParCOSNet, sensibly reduces the memory requirements, allowing predicting protein functions in few minutes on ordinary computers in networks with millions of nodes [15].

More recently, the candidate introduced also novel model of parametric Hopfield network, *HoMCat* (**Ho**pfield **M**ulti-**Cat**egory) [24], which can be seen as a generalization of *COSNet*. *HoMCat* allows to partition the input instances into multiple categories determined according to intrinsic properties the instances possess, independently from their membership to the class being predicted. Such an approach originated from the existence of many real-world problems in which instances can be naturally partitioned into categories, ranging from recommending systems to the multi-species gene function prediction, in which nodes in the graph can be partitioned according to the species they belong to. The model dedicates one parameter to each input category, to be learned in order to better fit the topological structure of that category in the graph. The adopted learning procedure is fast and scalable: its complexity indeed only linearly increases with the number of input categories. The application of *HoMCat* in the multi-species protein function prediction, where the

input graph is composed of genes belonging to different phylogenetically related species, achieved highly competitive results [24]. Furthermore, another variant of the model which automatically learns the partition in categories has been proposed [49].

6.1.2 Development of succinct classifiers and data structures

The ever growing need to efficiently store, retrieve and analyze massive datasets, originated by very different sources, is currently made more complex by the different requirements posed by users and applications. Such a new level of complexity cannot be handled properly by current data structures for big data problems.

In this context he developed techniques to compress Deep Neural Networks (DNN) and novel compressed storage for sparse and quantized matrices [3, 32, 33]. Applications of such techniques have led to promising results in the bio-medical applications in resource limited contexts [1]. In addition, in the domain of learned data structures, he preliminary investigated the role of classifiers in learned Bloom filters [31].

6.1.3 Automated integration of heterogeneous data sources

It is well known that the solutions to classification problems based on graph representation are strongly dependent on the graph topology itself. In several contexts, multiple and heterogeneous sources of information describing pairwise similarity relationships among nodes/instances are available; each source, represented through a dedicated graph, emphasizes only some characteristics/properties of nodes/instances, and can be highly predictive for some classification problems, and not predictive the other ones. Moreover, a single source sometimes covers solely a subset of the universe of instances; for these reasons the integration of heterogeneous data sources in a unique, composite graph, more reliable and with higher coverage, is a central and challenging problem in several research contexts, such as bioinformatics.

However, several methods proposed in the literature for graph integration completely neglect one of the main limitations affecting a large number of real context, namely the strong imbalance of node labels toward negatives in binary classification problems. Such characteristic of several real world problems, ranging from recommending systems to fraud detection, gene function prediction and disease-gene prioritization, is instead of paramount importance to assess the informativeness of each single graph/source for the classification problem to be handled.

In this area, the candidate studied a novel algorithm for heterogeneous graph integration, *UNIPred* (**Un**balance-aware **I**ntegration and **Pred**iction), which explicitly handle the labeling imbalance [25]. In particular, for every graph separately, the algorithm introduces a projection of nodes onto a bi-dimensional space, such that the position of projected points is a representative of the label imbalance at the corresponding node neighborhood [46, 51]. Projected points, which preserve their initial label, are the associated with an objective function imbalance-aware, whose optimization provides the index of informativeness for the related graph. The better the optimum, to more informative the graph/source. In order to assure scalability to large size graph, the optimization is performed through an approximated procedure, whose complexity is quasi-linear in the number of points.

The candidate is also working to the implementation of a web tool for *UNIPred*, with the aim of providing an easy-to-use interface allowing its usage even to researchers not familiar with computer programming.

6.1.4 Selection of negative examples in Positive-Unlabeled classification problems

Most machine learning algorithms for binary classification problems require reliable positive and negative examples to determine good quality solutions. Nevertheless, negative examples are just

rarely stored and annotated in public databases, unlike the positive ones. Classification problems where solely positive examples are well defined are known as *Positive-Unlabeled* (PU) learning problems. In this context, non positive instances are considered as unlabeled.

To this end, the candidate studied a novel algorithm for selecting among unlabeled examples those that can be considered as reliable negative examples. The method transforms the input graph into a feature matrix, in which each node is associated with a feature vector (point) depending on the classification problem and its topology in the graph, then positive points are clustered using a dynamic version of the fuzzy C means algorithm exploiting a suitable index in order to decide the optimal number of clusters to summarize data. Finally every negative point is assigned a score consisting in the maximum membership it has to the detected clusters of positive items. The low-ranked points according to this score are considered as negatively labeled.

He applied this negative selection procedure to real-world PU-problems by extending a classifier based on Hopfield networks [44], and a ranking algorithm based on weighted linear regression [21].

6.1.5 Multitask algorithms for hierarchical classification

This part of the candidate's research activity involves classification problems where different prediction tasks are related to each other and/or forming a hierarchy, and where each node/instance can be classified for more than one task at the same time (multilabel problems). It is well known that appropriately learning related tasks at the same time can improve and speed up learning, with regard to learning tasks independently. Indeed, the traditional approach is to learn related task so as to learn similar predictive models for similar tasks. The candidate studied also a different approach exploiting task dissimilarities rather than similarities and has shown that dissimilarity information enforces separation of rare class labels from frequent class labels, and for this reason is better suited for solving unbalanced classification problems [42]. In particular, the candidate developed two multitask extensions of the well know label propagation algorithm, the first one learning from similar tasks, the second one instead leveraging task dissimilarities, with promising results in a real-world application [20, 42].

When singletask algorithms instead of those multitask are adopted to infer predictions for tasks/classes structured as a hierarchy with parent-child relationships, the common problem arising is the violation of parent-child constraints, that is the same instance receives a higher membership to child classes than to parent ones. Several real-world taxonomies posses this distinctive feature, like the *Gene Ontology* (GO) for protein function prediction, the *Human Phenotype Ontology* for human abnormal phenotypes and so on.

To face this problem, the candidate also developed methods for the hierarchical correction of singletask predictions to force the respect of hierarchical parent-child constraints. In particular, tasks/nodes in the hierarchy are scanned by levels in two phases, bottom-up (from leaves to root nodes) and top-down, to modify the predictions for each instance separately regarding all the tasks and correct their constraint violations. The related preliminary results are promising [45].

Finally, another central issue in multitask learning is the adoption of reliable measures of task similarities. To this end, in collaboration with the Computer Science and Centre for Systems and Synthetic Biology of Royal Holloway of London, he contributed to the implementation of a web server to provide researchers of an efficient mean to compute and compare the main similarity measures between GO terms [27, 47], whose hierarchy forms a directed acyclic graph. The software is publicly available as both web server and JAVA standalone application[13].

[13] `http://www.paccanarolab.org/gossto`

6.1.6 Development of software machine learning libraries

The candidate always implemented in person the algorithms he developed, and made publicly available the software through public repositories, such as *GitHub* and *Bioconductor*, and also through Original Software Publications. Here the list:

- **COSNet**. **R** package downloadable from Bioconductor[14], main repository for Bioinformatics, and from GitHub repository[15], it implements the homonym classification algorithm [28], along with auxiliary utilities, including those to evaluate the generalization abilities of the method through k-fold cross validation procedures. The implementation adopt the **R** language for the routine functions, whereas the computational expensive parts are written in **C** language. *COSNet* is now also available as *Original Sofware Publication* on the *Neurocomputing* journal [23]. The methods is receiving around 1000 distinct downloads per year[16].
- **UNIPred**. Software library written in **C** and **R** languages, implementing the homonym method [25]. The software is publicly available at `http://frasca.di.unimi.it/Unipred.html`.
- **GOssTo**. JAVA standalone application and web-server[17] to compute semantic similarities between GO terms. The service is flexible, and allows the selection of GO terms, model organisms, and genes of interest. Furthermore, the application also allows to computed a gene-wise similarity based on their GO annotation profile.
- **RANKS**. Application note of the Bioinformatics journal describing the **R** package *RANKS* [22]. The package, also available at CRAN repository (`http://cran.r-project.org`), implements in **C** and **R** languages a ranking method of nodes in a graph based on kernel functions. The package also implements other state-of-the-art ranking methods and utilities to compare the performance of different methods.
- **ParCOSNet**. **C++** source code and executable program, publicy available at the GitHub platform[18], implementing the parallel version of COSNet [15], able to predict node labels in sparse graphs with million of nodes on standard computers equipped with GPU graphical devices.

6.2 Bioinformatics

The candidate's research activity in bioinformatics involved the study and application of machine learning algorithms for knowledge retrieval from bio-molecular data generated by high-throughput technologies. It is possible to distinguish the following lines:

6.2.A. Functional classification of genes and proteins

6.2.B. Integration of multiple and heterogeneous data sources for the gene/protein function prediction

6.2.C. Gene-disease prioritization for genetic human disorders

6.2.D. Multi-species protein function prediction

6.2.E. Epigenetic analysis of gene expression levels

[14] `https://www.bioconductor.org/packages/release/bioc/html/COSNet.html`
[15] `https://github.com/m1frasca/COSNet_GitHub`
[16] `http://bioconductor.org/packages/stats/bioc/COSNet/`
[17] `http://www.paccanarolab.org/gossto/`
[18] `https://github.com/anacleto63/ParCOSNet`

6.2.F. Biological motivated modelling of gene expression profiles

6.2.G. Drug repositioning methods to determine novel therapeutic applications of existing drugs

6.2.A. Functional classification of genes and proteins.
The functional classification of genes and their products, such as proteins, belonging to several species sequenced from the beginning of the 3th millennium, has become a central and challenging problem in bioinformatics. The various functions genes may possess are structured as hierarchies that describe their functional relationships: examples are the Gene Ontology, a DAG, and Functional Catalogue, a tree-based taxonomy created by the Munich Information Center for Protein Sequences (MIPS).

The problem is highly complex, with thousands of proteins and thousands of possible functions, and where each gene may possess multiple functions (terms in the hierarchy). Moreover, for most terms/functions, often just few or no genes are already annotated with that function, determining a strong imbalance between positive (those already associated with the function) and negative (the remaining ones) genes. A common habit of researchers in bioinformatics in the last years has been representing genes and their prior knowledge through gene networks/graphs, since biological functions often are extremely complex and involve multiple genes at the same time. Networks allow to represent in a natural way relationships among genes, and graph based methodologies are thereby required to face this problem.

In this context, the applicant developed network-based classifiers to predict functions of genes of several model organisms and for terms in the main biological taxonomies [22, 23, 26, 29]. In particular, he applied to this problem the imbalance-aware models developed in [28, 50], with results competitive with the state-of-the-art methods.

Since most biological taxonomies for gene functions introduce constraints between parent and child terms (functions), that is each gene associated with a child term must be associated with every parent term (*true path rule*), methods which predict each function independently from the other ones may violate the true path rule. To this end he applied the method for hierarchical correction developed in [45] to predict the Gene Ontology and the Human Phenotype Ontology terms, with promising preliminary results.

Finally, recently he also focused on the analysis of which protein features, extracted from protein-protein interaction networks, may be relevant for determining reliable negative proteins when learning machine learning models to predict the Gene Ontology functions [40]. Indeed, the Gene Ontology only stores positive annotations (positive examples), and all non-positive instances in principle can be used as negative to learn a classifier. This approach has been then extendend by including the most frequently used node centrality measures in graphs, with validation of the proteins features for both selecting negatives for and predicting the GO terms [17].

6.2.B. Integration of multiple and heterogeneous data sources for the gene/protein function prediction.
In the context of gene function prediction, various sources of data are available, such as gene expression profiles, genes-chemicals relationships, protein-protein physical and genetic interaction and so on, and each source in general covers a subset of genes and captures only some functional characteristics of genes. Moreover, each data source is in general predictive just for a subset of bio-molecular functions of genes, and may be not predictive for the other ones. For this reasons, integrating different data sources is a central problem in bioinformatics and particularly in the prediction of protein functions. Another important issue to be taken into account when integrating data sources is the strong imbalance between genes

which posses (positives) and those which do not posses (negatives) a given function in existing taxonomies. In this area, the candidate applied the algorithm proposed in [25] to a large number of species to integrate several gene networks and obtain more predictive networks for each term separately, and to infer functions for genes not associated to that term. The obtained results in the international challange MOUSEFUNC I for predicting functions of mouse proteins have been highly competitive.

6.2.C. Disease-gene prioritization for genetic human disorders.
Another challenging problem in computational biology and medicine is finding genes responsible or at least involved in human genetic disorders. Rarely, indeed, a genetic disorder is caused just by one gene, but often it is the results of the combined action of multiple biological pathways. Manually detecting causative genes is expensive and intractable, due to the large number of human genes (more that 20000) and genetic disorders (thousands). Accordingly, computational methods for this problem, often known as disease-gene prioritization algorithms, have been strongly required, at least to foster the research about some hot genes for specific diseases.

To face the disease-gene prioritization problem, the candidate developed a scalable algorithm, WGP (*Weighted Gene Prioritization*) to rank human genes according to a given disease, by also handling the disproportion between known causative and non causative genes: the higher the ranking, the stronger the clue the gene is involved in the disease etiology. The method performed among the top methods in a public benchmark comparison to predict the Medical Subject Headings (MeSH) disease terms [43]. An first extension of this method to embed a negative selection procedure improved the model accuracy [21], whereas a second and more promising extension, embedding in the model information from diseases with similar genetic patterns noticeably improved the prioritization abilities of the model, providing prioritization of disease-genes even for diseases without known associated genes [19].

6.2.D. Multi-species protein function prediction.
Many species in nature are phylogenetically correlated, that is they posses common ancestors in the evolutionary process. This means that homologous proteins, that is proteins coded in two different organisms by a gene inherited from a common ancestor species, may have similar functions. Thus, integrating proteins belonging to different species phylogenetically correlated in a unique network may improve the prediction of every individual species.

The *multi-species protein function prediction* (MFP) is thereby one of the most appropriate problems for the application of HoMCat, the method the candidate proposed in [25], where the input categories are the species composing the network. The results obtained in predicting proteins of *Danio rerio*, *Xenopus laevis*, and different species of bacteria have shown remarkable improvements with reference to results on single species networks. The candidate is also studying an extension to parallelize the method and make feasible its application to networks with hundreds of species.

Furthermore, other issues in MFP are the need of integrating information from several data banks and of appropriately visualize the predictions. With this aim, recently he published a position paper about a framework including all the moments of multi-species protein function prediction: data retrieval and processing from different data banks and organisms, their integration in a unique large-size network, the prediction of biomolecular functions of proteins, and the visualization of the obtained results [41].

6.2.E. Epigenetic analysis of gene expression levels.
Gene expression is a very complex process, which is finely regulated both at the genetic and epigenetic level. In particular, the latter, which involves the structure formed by DNA

wrapped around histones (chromatin), has been recently shown to be a key factor in determining gene expression levels. Hence building models to predict gene expression levels from epigenetic factors, such as histone modifications or transcription factors activities, is a key problem in bioinformatics.

During his research period at the Life Sciences Department of University of Milan the candidate focused on the study of a parametric model based on neural networks which ranks genes according to the information coded in the histone modifications occurring at genes [48]. The model has been validated in predicting gene expression levels of six human cell lines in a genome wide approach.

6.2.F. Biological motivated modelling of gene expression profiles.
The evaluation of gene selection methods is a challenging problem in bioinformatics. In most of cases, the genes associated with a specific phenotype are not a priori known, hence the evaluation of the effectiveness of gene selection methods is difficult and only partially performed by using classification algorithms. In this context, the candidate contributed to develop a mathematical model for generating biologically-plausible gene expression data to test gene selection methods. The method models expression profiles and expression signatures of a specific phenotype through positive Boolean functions, and it is able in generating biologically plausible gene expression data starting from their expression signatures and from the genes associated with the phenotype of interest [30].

6.2.G. Drug repositioning methods to determine novel therapeutic applications of existing drugs.
Producing novel drugs is an expensive process affected by an elevate risk of failure. In the last decade a paradigm of pharmacological research, known as *Drug repositioning*, has emerged, consisting in the research of potential applications of existing drugs for other therapeutic categories. Indeed, it can sensibly reduce costs and time for developing novel drugs, which usually require 10-15 years and investment even larger than one billion dollars.

Even in this case, information about drugs can be represented through drug networks, where connections embed information about the connected drugs: interaction with the same proteins, genes or diseases on which both drugs are effective, and so on. In the work [23] a network-based classifier has been applied to predict novel DrugBank therapeutic categories for more than a thousand of drugs. The results achieved are competitive with methods specifically designed for this problem.

Publications

Research actually submitted to international peer-reviewed conferences and journals

S.1

[1] Gliozzo, J., Marinò, G., Bonometti, A., **Frasca, M.**, Malchiodi, D.. Reducing the Complexity of Deep Learning Models for Medical Applications in Resource-limited Contexts: A Use Case. *Health Informatics Journal*, 2022.

International peer-reviewed journals

J.1

[2] Ferrè, L., Clarelli, F., Pignolet, B., Mascia, E., **Frasca, M.**, Santoro, S. Sorosina, M., Bucciarelli, F., Moiola, L., Martinelli, V., Comi, G., Liblau, R., Filippi, M., Valentini, G., Esposito, F. Combining clinical and genetic data to predict response to fingolimod treatment in Relapsing Remitting Multiple Sclerosis patients: a precision medicine approach. *Journal of Personalized Medicine*, 2023, 13(1):122. doi:10.3390/jpm13010122.

J.2

[3] G. Marinò; A. Petrini; D. Malchiodi and **M. Frasca**. Deep Neural Networks Compression: A Comparative Survey and Choice Recommendations. *Neurocomputing*, 2023, 520:152-170. doi:10.1016/j.neucom.2022.11.072.

J.3

[4] V. Quarato, S. D'Antona, P. Battista, R. Zupo, R: Sardone, I. Castiglioni, D. Porro, **M. Frasca**, C. Cava. Transcriptional profiling of hippocampus identifies network alterations in Alzheimer's Disease. *Applied Sciences*, 2022, 12(10), 5035. doi:10.3390/app12105035.

J.4

[5] G. Dal Santo, **M. Frasca**, G. Bertoli, I. Castiglioni, C. Cava. Identification of key miRNAs in prostate cancer progression based on miRNA-mRNA network construction. *Computational and Structural Biotechnology Journal*, Volume 20, 2022, 864-873. doi:10.1016/j.csbj.2022.02.002.

J.5

[6] Cava, C.; Pisati, M.; **Frasca, M.**; Castiglioni, I. Identification of Breast Cancer Subtype-Specific Biomarkers by Integrating Copy Number Alterations and Gene Expression Profiles. *Medicina*, 2021, 57, 261. doi:10.3390/medicina57030261.

J.6

[7] M. Notaro, **M. Frasca**, A. Petrini, J. Gliozzo, E. Casiraghi, PN Robinson, G. Valentini HEMDAG: a family of modular and scalable hierarchical ensemble methods to improve Gene Ontology term prediction. *Bioinformatics*, 37:23, 2021. doi:10.1093/bioinformatics/btab485.

J.7

[8] A. Esposito, E. Casiraghi, F. Chiaraviglio, A. Scarabelli,E. Stellato, G. Plensich, G. Lastella, L. Di Meglio, S. Fusco, E. Avola, A. Jachetti, C. Giannitto, D. Malchiodi, **M. Frasca**, A. Beheshti, PN Robinson, G. Valentini, L. Forzenigo, G. Carrafiello . Artificial intelligence in predicting clinical outcome in Covid-19 patients from clinical, biochemical and Chest X-Ray analysis. *Reports in Medical Imaging*, 2021:14 27–39. doi:10.2147/RMI.S292314.

J.8

[9] P. Perlasca, **M. Frasca**, C. T. Ba, J. Gliozzo, M. Notaro, M. Pennacchioni, G. Valentini, M. Mesiti. Multi-resolution visualization and analysis of biomolecular networks through hierarchical community detection algorithms and web-based graphical tools. *PLOS ONE* 2020,(12): e0244241. doi:10.1371/journal.pone.0244241.

J.9

[10] E. Casiraghi and D. Malchiodi and G. Trucco and **M. Frasca** and L. Cappelletti and T. Fontana and A. A. Esposito and E. Avola and A. Jachetti and J. Reese and A. Rizzi and P. N. Robinson and G. Valentini. Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments. *IEEE Access*, 8, 2020. ISSN: 2169-3536. doi:10.1109/ACCESS.2020.3034032.

J.10

[11] A. Petrini, M. Mesiti, M. Schubach, **M. Frasca**, D. Danis, M. Re, G. Grossi, L. Cappelletti, T. Castrignano', P.N. Robinson, G. Valentini. parSMURF, a High Performance Computing tool for the genome-wide detection of pathogenic variants. *GigaScience*, 9:5, 2020. ISSN: 2047-217X.
doi:10.1093/gigascience/giaa052.

J.11

[12] J. Gliozzo, P. Perlasca, M. Mesiti, A. Petrini, E. Casiraghi, **M. Frasca**, G. Grossi, M. Re, A. Paccanaro, and G. Valentini. Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction. *Scientific Reports*, 10, 3612, 2020. ISSN: 2045-2322. doi:10.1038/s41598-020-60235-82019.

J.12

[13] N. Zhou et al.. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20:244, 2019. ISSN: 1474-760X.
doi:10.1186/s13059-019-1835-8.

J.13

[14] P. Perlasca, **M. Frasca**, C.T. Ba, M. Notaro, A. Petrini, E. Casiraghi, G. Grossi, J. Ghilozzo, G. Valentini, M. Mesiti. UNIPred-Web: a Web Tool for the Integration and Visualization of Biomolecular Networks for Protein Function Prediction. *BMC Bioinformatics*, 20:422, 2019. ISSN:1471-2105.
doi:10.1186/s12859-019-2959-2.

J.14

[15] **M. Frasca**, G. Grossi, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, and G. Valentini A GPU-based algorithm for fast node label learning in large and unbalanced biomolecular networks. *BMC Bioinformatics*, 19 (Suppl 10):353, 2018. ISSN:1471-2105.
doi:10.1186/s12859-018-2301-4.

J.15

[16] E. Casiraghi, V. Huber, **M. Frasca**, M. Cossa, M. Tozzi, L. Rivoltini, B.E. Leone, A. Villa and B. Vergani A novel computational method for automatic segmentation, quantification and comparative analysis of immunohistochemically labeled tissue sections. *BMC Bioinformatics*, 19 (Suppl 10):357, 2018. ISSN:1471-2105.
doi:10.1186/s12859-018-2302-3.

J.16

[17] P. Boldi, **M. Frasca** and D. Malchiodi. Evaluating the Impact of Topological Protein Features on the Negative Examples Selection. *BMC Bioinformatics*, 19 (Suppl 14):417, 2018. ISSN:1471-2105.
doi:10.1186/s12859-018-2385-x. Impact factor: 2.213

J.17

[18] S. Vascon, **M. Frasca**, R. Tripodi, G. Valentini and M. Pelillo. Protein Function Prediction as a Graph-Transduction Game. *Pattern Recognition Letters*, 2018. In press. ISSN:0167-8655. doi:10.1016/j.patrec.2018.04.002. Impact factor: 1.995

J.18

[19] **M. Frasca**. Gene2DisCo: Gene to Disease Using Disease Commonalities. *Artificial Intelligence in Medicine*, 82:34–46, 2017. ISSN:0933-3657.
doi:https://doi.org/10.1016/j.artmed.2017.08.001. Impact factor: 2.879

J.19

[20] **M. Frasca** and N. Cesa Bianchi. Multitask protein function prediction through task dissimilarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017. ISSN:1545-5963.
In press. doi:10.1109/TCBB.2017.2684127. Impact factor : 2.428

J.20

[21] **M. Frasca** and D. Malchiodi. Exploiting negative sample selection for prioritizing candidate disease genes. *G*enomics and Computational Biology, 3(3) 47, 2017. ISSN:2365-7154.
doi:10.18547/gcb.2017.vol3.iss3.e47.

J.21

[22] G. Valentini, G. Armano, **M. Frasca**, J. Lin, M. Mesiti, and M. Re. RANKS: a flexible tool for node label ranking and classification in biological networks. *B*ioinformatics, 32(18):2872–2874, 2016. ISSN 1367-4803.
doi: http://dx.doi.org/10.1093/bioinformatics/btw235. Impact factor : 5.481

J.22

[23] **M. Frasca** and G. Valentini. COSNet: an R package for label prediction in unbalanced biological networks. *Neurocomputing*, 2016. In press. Available online from 29 April 2016. ISSN 0925-2312.
doi: http://dx.doi.org/10.1016/j.neucom.2015.11.096. Impact factor : 3.317

J.23

[24] **M. Frasca**, S. Bassis, and G. Valentini. Learning node labels with multi-category Hopfield networks. *Neural Computing and Applications*, 27(6):1677–1692, 2016. ISSN 1433-3058. doi: 10.1007/s00521-015-1965-1. Impact factor : 4.213

J.24

[25] **M. Frasca**, A. Bertoni, and G. Valentini. UNIPred: Unbalance-Aware Network Integration and Prediction of Protein Functions. *Journal of Computational Biology*, 22(12):1057–1074, 2015. ISSN 1066-5277.
doi: 10.1089/cmb.2014.0110. Impact factor : 1.191

J.25

[26] **M. Frasca**. Automated gene function prediction through gene multifunctionality in biological networks. *Neurocomputing*, 162:48 – 56, 2015. ISSN 0925-2312.
doi: 10.1016/j.neucom.2015.04.007. Impact factor : 3.317

J.26

[27] H. Caniza, A. E. Romero, S. Heron, H. Yang, A. Devoto, **M. Frasca**, M. Mesiti, G. Valentini, and A. Paccanaro. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, 30(15):2235–2236, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu144. Impact factor : 5.481

J.27

[28] **M. Frasca**, A. Bertoni, M. Re, and G. Valentini. A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, 43:84 – 98, 2013. ISSN 0893-6080.
doi: 10.1016/j.neunet.2013.01.021. Impact factor : 7.197

J.28

[29] **M. Frasca**, A. Bertoni, and G. Valentini. Regularized network-based algorithm for predicting gene functions with high-imbalanced data. *EMBnet.journal*, 18(15):41–42, 2012. ISSN 2226-6089. doi: 10.14806/ej.18.A.377

J.29

[30] M. Muselli, A. Bertoni, **M. Frasca**, A. Beghini, F. Ruffino, and G. Valentini. A mathematical model for the validation of gene selection methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(5):1385–1392, Sept 2011. ISSN 1545-5963. doi: 10.1109/TCBB.2010.83. Impact factor : 2.428

National and international peer-reviewed conferences

C.1

[31] Fumagalli, G.; Raimondi, D.; Giancarlo, R.; Malchiodi, D. and **Frasca, M**. On the Choice of General Purpose Classifiers in Learned Bloom Filters: An Initial Analysis Within Basic Filters. *In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - ICPRAM22,*, pages 675-682. doi:10.5220/0010889000003122

C.2

[32] G. Marinò, G. Ghidoli, **M. Frasca**, D. Malchiodi. Reproducing the sparse Huffman Address Map compression for deep neural networks. *3th International workshop on Reproducible Research in Pattern Recognition (RRPR' 21)*. Lecture Notes in Computer Science, vol 12636. Springer, Cham. 2020. doi:10.1007/978-3-030-76423-4_12

C.3

[33] G. Marinò, G. Ghidoli, **M. Frasca**, D. Malchiodi. Compression strategies and space-conscious representations for deep neural networks. *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 9835-9842. doi:10.1109/ICPR48806.2021.9412209.

C.4

[34] **M. Frasca**, G. Grossi, G. Valentini. Multitask Hopfield Networks. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML19), 2019. LNCS, vol 11907. Springer, Cham, 2020 doi:10.1007/978-3-030-46147-8_21

C.5

[35] C. T. Ba, E. Casiraghi, **M. Frasca**, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, M. Re and G. Valentini. A Graphical Tool for the Exploration and Visual Analysis of Biomolecular Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. LNCS, vol 11925. Springer, Cham. doi:10.1007/978-3-030-34585-3_8

C.6

[36] **M. Frasca**, M. Sepehri, A. Petrini, G. Grossi and G. Valentini. Committee-based Active Learning to Select Negative Examples for Predicting Protein Functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. LNCS, vol 11925. Springer, Cham. doi:10.1007/978-3-030-34585-3_7

C.7

[37] M. Sepehri, **M. Frasca**. Analysis of Novel Annotations in the Gene Ontology for Boosting the Selection of Negative Examples. *In Proceedings of the 9th International Conference on Biomedical Engineering and Technology (ICBET' 19)*, 2019. ACM, 294–301. doi:10.1145/3326172.3326228

C.8

[38] **M. Frasca**, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi and M. Andrade-Navarro. Disease–Genes must Guide Data Source Integration in the Gene Prioritization Process. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, in press. ISSN:0302-9743.

C.9

[39] M. Notaro, M. Schubach, **M. Frasca**, M. Mesiti, P.N. Robinson and G. Valentini. Ensembling Descendant Term Classifiers to Improve Gene–Abnormal Phenotype Prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, in press. ISSN:0302-9743.

C.10

[40] **M. Frasca**, F. Lipreri and D. Malchiodi. Analysis of Informative Features for Negative Selection in Protein Function Prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10209 LNCS, 267-276, 2017. doi: 10.1007/978-3-319-56154-7_25. ISSN:0302-9743.

C.11

[41] G. Perlasca, P. Valentini, **M. Frasca**, and M. Mesiti. Multi-species protein function prediction: Towards web-based visual analytics. In *18th International Conference on Information Integration and Web-based Applications & Services(iiWAS2016), 28–30 Nov, Singapore.*, pages 1–5. ACM 2016. ISBN 978-1-4503-4807-2.

C.12

[42] **M. Frasca** and N. Cesa Bianchi. Multi-task label propagation with dissimilarity measures. In *15th International Workshop on Data Mining in Bioinformatics (BIOKDD 16), 14 Aug, San Francisco, CA, USA.*, pages 1–8, 2016. URL : `http://home.biokdd.org/biokdd16/`.

C.13

[43] **M. Frasca** and S. Bassis. Gene-Disease Prioritization Through Cost-Sensitive Graph-Based Methodologies, volume 9656 of *Lecture Notes in Computer Science*, pages 739–751. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-31744-1_64.
ISBN 978-3-319-31744-1.

C.14

[44] **M. Frasca** and D. Malchiodi. Selection of negative examples for node label prediction through fuzzy clustering techniques. In *Advances in Neural Networks. WIRN 2015. Smart Innovation, Systems and Technologies*, pages , vol 54:67–76, Springer, Cham. doi: 10.1007/978-3-319-33747-0_7.
ISBN:978-3-319-33746-3.

C.15

[45] P. N. Robinson, **M. Frasca**, S. Köhler, M. Notaro, M. Re, and G. Valentini. *A Hierarchical Ensemble Method for DAG-Structured Taxonomies*, *Multiple Classifier Systems. MCS 2015. Lecture Notes in Computer Science*, vol 9132 pages 15–26, 2015. Springer International Publishing, Cham. doi: 10.1007/978-3-319-20248-8_2.
ISBN 978-3-319-20248-8.

C.16

[46] **M. Frasca**, A. Bertoni, and G. Valentini. An unbalance-aware network integration method for gene function prediction. In *MLSB 2013 - Machine Learning for Systems Biology - Berlin, July 19-20*, Berlin, Germany, 2013.

C.17

[47] H.V. Caniza, A.E. Romero, S. Heron, H. Yang, **M. Frasca**, M. Mesiti, G. Valentini, and A. Paccanaro. Gossto & gosstoweb: user-friendly tools for calculating semantic similarities on the gene ontology. In *Bio-Ontologies SIG 2013 - ISMB 2013*, Berlin, Germany, 2013.

C.18

[48] **M. Frasca** and G. Pavesi. A neural network based algorithm for gene expression prediction from chromatin structure. In *The 2013 International Joint Conference on Neural Networks (IJCNN), 4-9 Aug, Dallas Texas,*, pages 1–8. IEEE, 2013. doi: 10.1109/IJCNN.2013.6706954. ISBN 978-1-4673-6128-6.

C.19

[49] **M. Frasca**, A. Bertoni, and A. Sion. A Neural Procedure for Gene Function Prediction. *Neural Nets and Surroundings. Smart Innovation, Systems and Technologies*, vol 19, pages 179–188, 2013. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-35467-0_19. ISBN 978-3-642-35467-0.

C.20

[50] A. Bertoni, **M. Frasca**, and G. Valentini. COSNet: A Cost Sensitive Neural Network for Semi-supervised Learning in Graphs. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science*, vol 6911, pages 219–234, 2011. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-23780-5_24. ISBN 978-3-642-23780-5.

C.21

[51] A. Bertoni, **M. Frasca**, and G. Valentini. An efficient supervised method to integrate multiple biological networks. In *BITS 2011, Bioinformatics Italian Society Annual Meeting, June 20-22*, Pisa, Italy, 2011.

C.22

[52] A. Bertoni, **M. Frasca**, G. Grossi, and G. Valentini. *Learning functional linkage networks with a cost-sensitive approach*, volume 226 of *Frontiers in Artificial Intelligence and Applications*, pages 52–61. IOS Press, 2010. doi: 10.3233/978-1-60750-692-8-52.

Conferences without proceedings

M. Frasca, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi and M. Andrade-Navarro. Disease–Genes must Guide Data Source Integration in the Gene Prioritization Process. *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, 2017, 7-9 Settembre, Cagliari, Italy.

M. Notaro, M. Schubach, **M. Frasca**, M. Mesiti, P.N. Robinson and G. Valentini. Ensembling Descendant Term Classifiers to Improve Gene–Abnormal Phenotype Prediction. *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, 2017, 7-9 settembre, Cagliari, Italy.

M. Frasca. *Selection of Negatives in Hopfield Networks*, *International Workshop on Dynamics of Multi-Level Systems (DYMULT) 2015, Max Planck Institute for the Physics of Complex Systems*, Dresden, 2015.